EP 3260 Homework 1 Group-2

1.1 Sol: Given that differentiable function f is μ-strongly convex iff $\forall x_1, x_2 \in X$, $\mu > 0$

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2 \quad -(1)$$

Proof 1: To be proved

$$\nabla^2 f(x) \geq \mu I_d \;, \quad \forall x \in X$$

Let us consider (1) & $x_1 = x_1$ & $x_2 = x$

$$\Rightarrow f(x) \geq f(x_1) + \nabla f(x_1)^T (x - x_1) + \frac{\mu}{2} \|x - x_1\|^2$$

apply gradient

$$\Rightarrow \nabla f(x) \geq \nabla f(x_1) + \nabla f(x_1)^T \cdot \nabla (x - x_1) + \frac{\mu}{2} \cdot \nabla \|x - x_1\|^2$$

$$\Rightarrow \nabla f(x) \geq 0 + \nabla f(x_1)^T \cdot (1 - 0) + \frac{\mu}{2} \cdot 2 \cdot \|x - x_1\|$$

Apply gradient again

$$\Rightarrow \nabla^2 f(x) \geq \nabla \left(\nabla f(x_1)^T \right) + \mu \cdot \nabla \|x - x_1\|$$

$$\Rightarrow \nabla^2 f(x) \geq 0 + \mu I_d$$

Hence proved.

Proof 2: To be proved

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$

If f is strongly convex then it is also true that

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \|x_1 - x_2\|^2 \quad -(2)$$

By adding (1) & (2)

$$\Rightarrow f(x_1) + f(x_2) \geq f(x_1) + f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_1 - x_2\|^2 + \frac{\mu}{2} \|x_2 - x_1\|^2$$

$$\Rightarrow \quad 0 \geq (\nabla f(x_2) - \nabla f(x_1))^T (x_1 - x_2) + \mu \cdot \|x_2 - x_1\|^2$$

$$\Rightarrow \quad (\nabla f(x_2) - \nabla f(x_1))^T \cdot (x_2 - x_1) \geq \mu \cdot \|x_2 - x_1\|^2 \quad -(3)$$

Hence proved.

<u>proof 3.a</u>: To be proved

$$f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2, \forall x$$

Consider (1) & let $x_1 = x^*$, $x_2 = x$

$$\Rightarrow \quad f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{\mu}{2} \|x - x^*\|^2$$

$$\Rightarrow \quad f(x) - f(x^*) \geq 0 + \frac{\mu}{2} \|x - x^*\|^2 \quad -(4)$$

Consider (3) let $x_1 = x^*$, $x_2 = x$

$$\Rightarrow (\nabla f(x) - \cancel{\nabla f(x^*)}^{0})^T (x - x^*) \geq \mu \|x - x^*\|^2$$

$$\Rightarrow \quad \nabla f(x)^T \cdot (x - x^*) \geq \mu \|x - x^*\|^2$$

Apply norm on both sides

$$\Rightarrow \quad \|\nabla f(x)\| \cdot \|x - x^*\| \geq \mu \|x - x^*\|^2$$

$$\Rightarrow \quad \|\nabla f(x)\| \geq \mu \cdot \|x - x^*\|$$

$$\Rightarrow \quad \|x - x^*\| \leq \frac{\|\nabla f(x)\|}{\mu} \quad -(5)$$

From (4) & (5) we get

$$f(x) - f(x^*) \leq \frac{\mu}{2} \cdot \frac{\|\nabla f(x)\|^2}{\mu^2}$$

$$\therefore f(x) - f(x^*) \leq \frac{1}{2\mu} \cdot \|\nabla f(x)\|^2$$

Hence proved.

Proof 3.b : To be proved

$$\|x_2 - x_1\| \leq \frac{1}{\mu} \cdot \|\nabla f(x_2) - \nabla f(x_1)\|$$

Let us consider ③

$$\Rightarrow (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|^2$$

apply norm on both sides

$$\Rightarrow \|\nabla f(x_2) - \nabla f(x_1)\| \cdot \|x_2 - x_1\| \geq \mu \cdot \|x_2 - x_1\|^2$$

$$\Rightarrow \|\nabla f(x_2) - \nabla f(x_1)\| \geq \mu \cdot \|x_2 - x_1\|$$

$$\Rightarrow \|x_2 - x_1\| \leq \frac{1}{\mu} \cdot \|\nabla f(x_2) - \nabla f(x_1)\| \quad - ⑥$$

hence proved.

Proof 3.c : To be proved

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \leq \frac{1}{\mu} \cdot \|\nabla f(x_2) - \nabla f(x_1)\|^2$$

We know that

$$\langle \bar{a}, \bar{b} \rangle \leq \|\bar{a}\| \cdot \|\bar{b}\|$$

$$\Rightarrow (\nabla f(x_2) - \nabla f(x_1))^T \cdot (x_2 - x_1) \leq \|\nabla f(x_2) - \nabla f(x_1)\| \cdot \|x_2 - x_1\|$$

from ⑥

$$\Rightarrow (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \leq \frac{\|\nabla f(x_2) - \nabla f(x_1)\|^2}{\mu}$$

Hence proved.

Proof 3.d: To be proved

$f(x) + r(x)$ is strongly convex for any convex f and strongly convex r

We know that if f is convex then

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T \cdot (x_2 - x_1) \quad -(7)$$

and if $r(x)$ is strongly convex then

$$r(x_2) \geq r(x_1) + \nabla r(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2 \quad -(8)$$

Now, let $g(x) = f(x) + r(x)$

In order for $g(x)$ to be strongly convex it should satisfy following

$$g(x_2) \geq g(x_1) + \nabla g(x_1)^T \cdot (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2$$

To validate this, let us add (7) & (8)

$$f(x_2) + r(x_2) \geq f(x_1) + r(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \nabla r(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2$$

$$f(x_2) + r(x_2) \geq f(x_1) + r(x_1) + \left(\nabla f(x_1) + \nabla r(x_1)\right)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2$$

$$\Rightarrow \quad g(x_2) \geq g(x_1) + \nabla g(x_1)^T \cdot (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2$$

Hence proved that $g(x)$ is strongly convex

1.2 Sol: Given that f is L-Smooth $\forall x_1, x_2$

$$\|\nabla f(x_2) - \nabla f(x_1)\| \leq L \|x_2 - x_1\|$$

Proof @: To be proved

$$f(x_2) \leq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2} \|x_2 - x_1\|^2$$

Let us consider

$$\left| f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \right|$$

$$= \left| \int_0^1 \nabla f(x_1 + \theta(x_2 - x_1))^T (x_2 - x_1)\, d\theta - \nabla f(x_1)^T (x_2 - x_1) \right|$$

$$\leq \int_0^1 \|\nabla f(x_1 + \theta(x_2 - x_1)) - \nabla f(x_1)\| \cdot \|x_2 - x_1\|\, d\theta$$

$$\leq \int_0^1 \theta \cdot L \cdot \|x_2 - x_1\| \cdot \|x_2 - x_1\|\, d\theta$$

$$\leq \frac{L}{2} \|x_2 - x_1\|^2$$

Hence proved.

Proof (b): To be proved

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{1}{2L} \cdot \|\nabla f(x_2) - \nabla f(x_1)\|^2$$

Let us consider

$$z = x_2 - \frac{1}{L} (\nabla f(x_2) - \nabla f(x_1))$$

$$f(x_1) - f(x_2) = f(x_1) - f(z) + f(z) - f(x_2)$$

$$\leq \nabla f(x_1)^T (x_1 - z) + \nabla f(x_2)^T (z - x_2) + \frac{L}{2} \|z - x_2\|^2$$

$$= \nabla f(x_1)^T (x_1 - x_2) + (\nabla f(x_1) - \nabla f(x_2))^T (x_2 - z)$$

$$+ \frac{1}{2L} \cdot \|\nabla f(x_1) - \nabla f(x_2)\|^2$$

$$= \nabla f(x_1)^T (x_1 - x_2) - \frac{1}{2L} \cdot \|\nabla f(x_1) - \nabla f(x_2)\|^2$$

$$\Rightarrow f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{1}{2L} \| \nabla f(x_2) - \nabla f(x_1) \|^2 \quad \rightarrow (1)$$

Hence proved.

Proof (c): To be proved

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \frac{1}{L} \| \nabla f(x_2) - \nabla f(x_1) \|^2$$

Let us consider (1) & interchange $x_1 \Leftrightarrow x_2$

$$\Rightarrow f(x_1) \geq f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \frac{1}{2L} \| \nabla f(x_1) - \nabla f(x_2) \|^2 \quad \rightarrow (2)$$

By adding (1) & (2)

$$\Rightarrow 0 \geq \nabla f(x_1)^T (x_2 - x_1) + \nabla f(x_2)^T (x_1 - x_2) + \frac{1}{L} \| \nabla f(x_2) - \nabla f(x_1) \|^2$$

$$\Rightarrow (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \frac{1}{L} \| \nabla f(x_2) - \nabla f(x_1) \|^2$$

Hence proved.

Sol 1.3:

(a) Sublinear convergence rate:

A sequence of update x_k is said to have a sublinear convergence rate if the norm of current appx. and optimal solution decreases at a slower rate than linear as the no. of iterations increase

$$\|x_k - x^*\| \leq \frac{c}{k} \quad \text{where} \quad c > 0$$

Benefits: Slower than linear but can lead to an accurate appx. with large k.

Eg:- Gradient descent with constant step size has sublinear convergence rate

(b) Linear CR:

A sequence of update x_k is said to have a linear convergence rate if the norm of current appx. and optimal solution decreases at a linear rate as no. of iterations increase.

$$\|x_k - x^*\| \leq \frac{c}{k}, \quad (c > 0)$$

Benefits: Faster than sublinear, but still relatively slow compared to super linear & Quadratic convergence. Can lead to accurate appx. with reasonable k.

Eg: Newton's method for optmizing a convex function has a linear convergence rate.

(c) Super Linear CR:

A sequence of update x_k is said to have a superlinear convergence rate if the norm of current appx. and optimal solution decreases at a faster rate than Linear as k increases.

$$\| x_k - x^* \| = \frac{c}{k^2} \quad , (c > 0)$$

Benefits: Faster than Linear convergence & can lead to accurate appx. with smaller no. of iterations. This is useful when the computational resources are limited.

Eg: Conjugate gradient method for optimizing a quadratic function has a superlinear CR.

(d) Quadratic CR:

A sequence of update x_k is said to have a Quadratic convergence rate if the norm of current appx. and optimal solution decreases at rate proportional to the square no. of iterations.

$$\| x_k - x^* \| \leq \frac{c}{k^2} \quad , (c > 0)$$

Benefits: Faster than all convergence rates, which leads to the accurate appx. with very small no. of iterations. This is useful when computational resources are limited.

Eg:- The Newton-Raphson method for optimizing a smooth & strongly convex function has a Quadratic CR.

1.4 Sol:

(a) When $N=1000$, the minimization problem can be solved using various optimization algorithms such as Gradient descent (GD), GD with line search, Newton's method, conjugate gradient etc. The choice of optimization would depend on the specifics of problem, such as structure of f and A, the desired convergence rate and amount of computation that can be done.

In case of strong convexity & smoothness of f, using a second-order optimization algorithm such as Newton's method would be faster in terms of convergence rate compared to the first order algorithms such as GD. However, these 2nd order optimization algorithms typically require the computation and storage of the Hessian which can be computationally expensive for large N.

(b) When $N = 10^9$, it becomes infeasible to use traditional methods like GD or Newton's method that require computation and storage of the gradient or Hessian matrix which would be of size $N \times N$. Instead, methods like stochastic GD variants like mini-batch SGD can be used.

(c) Using Newton's method with $N = 10^9$ can be computationally infeasible due to the computation and storage requirements for the Hessian matrix which would be of order $N \times N$, making it impractical for $N = 10^9$.

An efficient method for computing hessian matrix for probability simplex constraint with $P = 1$ & $b = 1$ would be to use the Limited-memory BFGS (L-BFGS) method, which is an optimization algorithm that approximates the inverse of Hessian matrix using a limited memory.

Extending this method to $1 \leq P \leq N$ would involve adjusting the memory requirements of L-BFGS method to accomodate the increased size of the Hessian matrix. Techniques like parallel computing, distributed computing improve the computational efficiency of the opt. problem.

(d) When adding a twice differentiable regularization term, $r(x)$ to the objective function, the optimization problem becomes

$$\min \quad f(x) + r(x)$$

where, $f(x)$ is original objective function & $r(x)$ is the regularization term. The purpose of adding regularization term is to prevent over fitting & improve performance of the model.

For $N = 1000$, we can still solve using above algorithms but with an additional step of computing gradient of $r(x)$, i.e now, $\nabla r(x)$ is added to $\nabla f(x)$. For $N = 10^9$, the algorithms like L-BFGS of SGD may be more appropriate

The common regularization ($r(x)$) terms include L1 regularization (LASSO) and the L2 regularization (Ridge regression)

We let $\varphi(x) = f(x) - \frac{\mu}{2}\|x\|^2$. $\varphi(x)$ is convex iff $f(x)$ is μ-strongly convex.

We can see $\varphi(x)$ is $(L-\mu)$-smooth, by using Lemma 3.4 in [Bubeck]. Moreove, by using Lemma 3.5, we have.

$$(\nabla\varphi(x) - \nabla\varphi(y))^T (x-y) \geq \frac{1}{L-\mu} \|\nabla\varphi(x) - \nabla\varphi(y)\|^2$$

We notes $\nabla\varphi(x) = \nabla\left(f(x) - \frac{\mu}{2}(x^T x)\right) = \nabla f(x) - \mu x$

$$(\nabla f(x) - \nabla f(y))^T (x-y) - \mu\|x-y\|_2^2 \geq \frac{1}{L-\mu}\Big[\|\nabla f(x) - \nabla f(y)\|^2$$

$$- \mu(\nabla f(x) - \nabla f(y))^T (x-y) - \mu(x-y)^T(\nabla f(x) - \nabla f(y))$$

$$+ \mu^2\|x-y\|^2$$

$$(L-\mu)\left(\nabla f(x)-\nabla f(y)\right)^T(x-y)+2\mu\left(\nabla f(x)-\nabla f(y)\right)^T(x-y)\geq$$

$$(L-\mu)\mu\|x-y\|_2^2+\|\nabla f(x)-\nabla f(y)\|_2^2+\mu^2\|x-y\|_2^2$$

By grouping terms,

$$(L+\mu)\left(\nabla f(x)-\nabla f(y)\right)^T(x-y)\geq\|\nabla f(x)-\nabla f(y)\|_2^2+L\mu\|x-y\|_2^2$$

Therefore

$$\left(\nabla f(x)-\nabla f(y)\right)^T(x-y)\geq\frac{1}{L+\mu}\|\nabla f(x)-\nabla f(y)\|_2^2+\frac{L\mu}{L+\mu}\|x-y\|_2^2$$